Navigating and Analyzing Large Economic Datasets*part2*

Xiaoyi Zhao

Virginia Tech Data Science for the Public Good Program

xiaoyiz@vt.edu

VIRGINIA TECH

Table of Contents

Checklist for Thoughtful Cleaning

Before You Clean, Ask Yourself…

- What am I trying to measure?
- What assumptions am I making when I clean my data?

What do you need before you clean your data

- **A closely related research paper** for guidance on variable definitions and modeling choices
- A clear list of variables:
  - Dependent variable(s)
  - Key explanatory variables and controls
- **The dataset's codebook** to understand variable labels, value codes, and formats

What are the things you need to look into when we clean

- Refer to your benchmark paper:
  - Check how similar studies define and clean variables.
  - Take a look at the assumptions they're working with, and think about how your own research approach might differ
- Adjustments or normalizations:
  - Are values adjusted for inflation? Or are they adjusted for seasonality?

What are the things you need to look into when we clean

- Units and scale:
  - For example, is income measured in dollars, thousands, or percentiles?
- Model structure matters:
  - Are you planning to use a linear model? Make sure your variable transformations align with that choice.

Steps

- Convert the format of the dataset (if we cannot directly work with the format you downloaded)
- Remove allocated or unreliable observations
- Create and rename variables
- Handle missing values
- Merge with other datasets

Tips

- Name the variables descriptively so someone else (or future you) can follow your work.
  - good example: age_female
  - bad example: age_2
- Keep track of how you transform raw variables (e.g., categorizing income levels or recoding region)

Tips

- Being consistent with file name and variable names
  - For example, for the data acs_2000, set the name acs_2000_i(after clean the allocated), acs_2000_pre(after generate the variables), and acs_2000_final
- Validate at each step
- Separate your data cleaning steps into modular files, and manage the full workflow through a master script—so anyone reviewing your work only needs to run that one file.

Thank You for listening!

Questions and Discussion.